**Colorado Technical**

 **Course:** MATH366 – Probability and Statistics

#### Unit 3 Part 06 Readings: Correlation and Time Series

**Correlation**

a relationship between two variables (x,y)

correlation coefficient - a mathematical measure of the strength and direction

 of a linear relationship between two variables

symbolized “r”

$$r=\frac{\left(Σxy\right)-(Σx)(Σy)}{\sqrt{\left[\left(Σx^{2}\right)-\left(Σx\right)^{2}\right][\left(Σy^{2}\right)-(Σx)^{2}]}}$$

positive correlation - as one variable increases,

**Interpreting Correlation Values**

–1.00 A perfect negative linear relationship

–0.70 A strong negative linear relationship

–0.50 A moderate negative relationship

–0.30 A weak negative linear relationship

 0.00 No linear relationship

+0.30 A weak positive linear relationship

+0.50 A moderate positive relationship

+0.70 A strong positive linear relationship

+1.00 A perfect positive linear relationship

the other also increases

negative correlation - as one variable increases,

the other decreases

zero correlation - no LINEAR relationship - may

be nonlinear

**correlation does not imply causation**

 RSQ = r2  the percent of the variability in *x*

that can be explained by *y* (or vice versa)

**Regression**

 the technique of fitting a linear equation to data

 regression is used to: estimate or predict the value *y* (the response variable) for a

given *x* (the predictor variable)

 **equation of a regression line:**

y = b + mx

m is the slope of the equation

b is the y-intercept

y is the calculated y-value for a given x-value

 RSQ measures how accurately a regression line predicts the data

**Using regression equations**

 regression is used not only to fit data but to forecast (predict) observations -

especially for a time series

 for a time series, “time” is always the “x” (independent) variable

 the method of fitting the line is least squares (minimum vertical distance)

**Calculating values using a regression equation:** y = b + mx

 Suppose you had:

value = $15,000 - $500(age)

 To find the value if the age is 10, plug that value in to the “age” slot in the equation:

value = $15,000 - $500(10)

 = $15,000 -$5,000

 = $10,000

**Interpreting regression equations:**

 For the equation above:

 The value when the item is new (age 0) is b = $15,000

 The value decreases m = $500 each year the item ages

#### Time Series

data that change over time

the goal is often to create an accurate forecast of future data values

spikey data are called "**volatile**"

it is extremely hard to forecast a volatile dataset

information in a spreadsheet about the data is called **metadata**

**Forecasting vs estimating using regression equations:**

forecasts use an x value beyond the original range of the data

for example, if the value equation above was calculated using data with ages

ranging from 0 to 15 years, plugging in any value of “age” from 0 to 15 would

be an estimate

If you plug in a value of 20 for “age”, that would be a “forecast” because an age of 20

was not in the original data used to calculate the equation

estimates have a tendency to be inexact (especially if the correlation is low)

forecasts tend be poorer and to get worse and worse as the x values move away

from the values used to calculate the equation